Machine Learning - Intro

Aarti Singh & Geoff Gordon

Machine Learning 10-701
Feb 1, 2021

Teaching team

Instructors:

Geoff

Aarti

Education Associate:

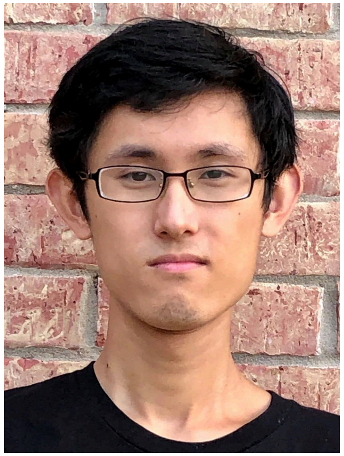
Brynn

Admin:

Mary

Teaching team

TAs:



Jeffrey H

Stefani

Yuchen

Arundhati

Jeffrey T

Logistics

Lectures: Mon, Wed 4:00-5:20 pm Remote

Recitations: Fri 4:00-5:20 pm Remote

Office hours:

Day	Time	Location	Staff
Mon			
Tues			
Wed			
Thurs	9-10 am	Remote	Aarti
Sat	1:30-2:50 pm	GHC 4401, Remote	

Zoom links: Canvas

Lectures will be recorded. Strictly for your use only.

Breakout rooms and Office hours will NOT be recorded.

Logistics

In case of technical issues during lecture:

please try logging back in

if issues remains > 5 mins, we will send an email when resolved.

If not resolved, recorded lecture + extra office hours

Webpage: https://www.cs.cmu.edu/~aarti/Class/10701_Spring21
Syllabus, policies, schedule of lectures, recitations, office hours, slides, reading material, homeworks, …

Piazza: http://piazza.com/cmu/spring2021/10701
announcements, questions for Teaching team, discussion forum for students

Homework submission: Gradescope

Grades: Canvas

Expectations

- Remote sessions (lecture, recitation, breakout room, office hours)
  - Turn on video if possible, especially for breakout rooms and office hours
  - Keep yourself muted unless asking or responding to questions

  Interact!
  - Ask questions in class by raising hand or via Zoom chat
  - Respond to questions in class by raising hand or via Zoom chat
  - Anonymous polls on Zoom
  - Breakout rooms

- In-person Office hours (Sat after Feb 15)
  - Only attend OH you are assigned to
  - Follow CMU guidance on masks, distancing and physical space, etc.

Recitations

- Strongly recommended
  - Brush up pre-requisites
  - Hands-on exercises
  - Review material (difficult topics, clear misunderstandings, extra new topics, HW and exam solutions)
  - Ask questions

- 1st Probability Review - **FRIDAY**

  by Jeffrey

  Fri Feb 5 4:00-5:20 pm Remote

Grading

- Grading
  - 5 homework assignments (5 x 10% = 50%)
  - 2 depth exercise (2 x 10% = 20%)
  - 1 midterm, 1 final: (12+15 = 27%)
    midterm - Mar 12 during recitation
  - Participation (3%)

- Late days
  - total 7 across homeworks, no more than 2 per HW
  - 50% credit for 24 hrs after late days
  - late days are for unforeseen situations (interviews, conference, etc.), do NOT include them in your plan

Homeworks & QnAs

- Collaboration
  - You may **discuss** the questions
  - Each student writes their own answers, without copying from discussion notes or ongoing conversations
  - Each student must write their own code for the programming part
  - **Don't search for answers on the web, Google, previous years' homeworks, etc.**
    - please ask us if you are not sure if you can use a particular reference
    - list resources used (references, discussants) on top of submitted homework
- Homeworks are hard, start early ☺
- Due on gradescope

Waitlist + Audits + Pass/Fail

- Waitlist

  we'll let everyone in

  keep attending lectures, recitations and office hours virtually and doing HW

- Audits and Pass/Fail

  Audits allowed (with some requirement)

  Pass/Fail allowed

About the course

- Machine Learning **Algorithms, Theory, Principles and Applications**
  - Classification: Naïve Bayes, Logistic Regression, Neural Networks, Support Vector Machines, k-NN, Decision Trees, …
  - Regression: Linear regression, Kernel regression, Nonparametric regression, …
  - Unsupervised methods: Kernel density estimation, mixture models, clustering, PCA, …
  - Graphical models, Hidden Markov Models, Reinforcement learning
  - Core concepts: Probability, Optimization, Theory, Model selection, overfitting, bias-variance tradeoffs, Fairness …

- See **tentative** lecture schedule on webpage – MAY CHANGE
- Material: Class slides + Reading material

Recommended textbooks

- Textbooks (Recommended, not required):

  Pattern Recognition and Machine Learning, Christopher Bishop (available online)

  Machine Learning: A probabilistic perspective, Kevin Murphy

  (available online)

  Machine Learning, Tom Mitchell

  The elements of statistical learning: Data mining, inference and prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman

Pre-requisites

- **Assume mathematical maturity**
  - Basic Probability and Statistics

    Probability distributions – discrete and continuous, Mean, Variance, Conditional probabilities, Bayes rule, Central limit theorem…
  - Programming (python) and principles of computing
  - Multivariate Calculus

    Derivatives, integrals of multi-variate functions
  - Linear Algebra

    Matrix inversions, eigendecomposition, …

- **Tutorial videos**
  - Probability, Calculus, Functional Analysis, SVD

    https://www.youtube.com/channel/UC7gOYDYEgXG1yIH_rc2LgOw/playlists
  - Linear Algebra

    http://www.cs.cmu.edu/~zkolter/course/linalg/index.html

- **Self-assessment test** on webpage

Fun begins!

- Why is ML trending?

applicability

large datasets < large # data points

large # features ← characteristics of data points

computing power, GPU, cloud

automate

low entry barriers

rule based algorithms don't work

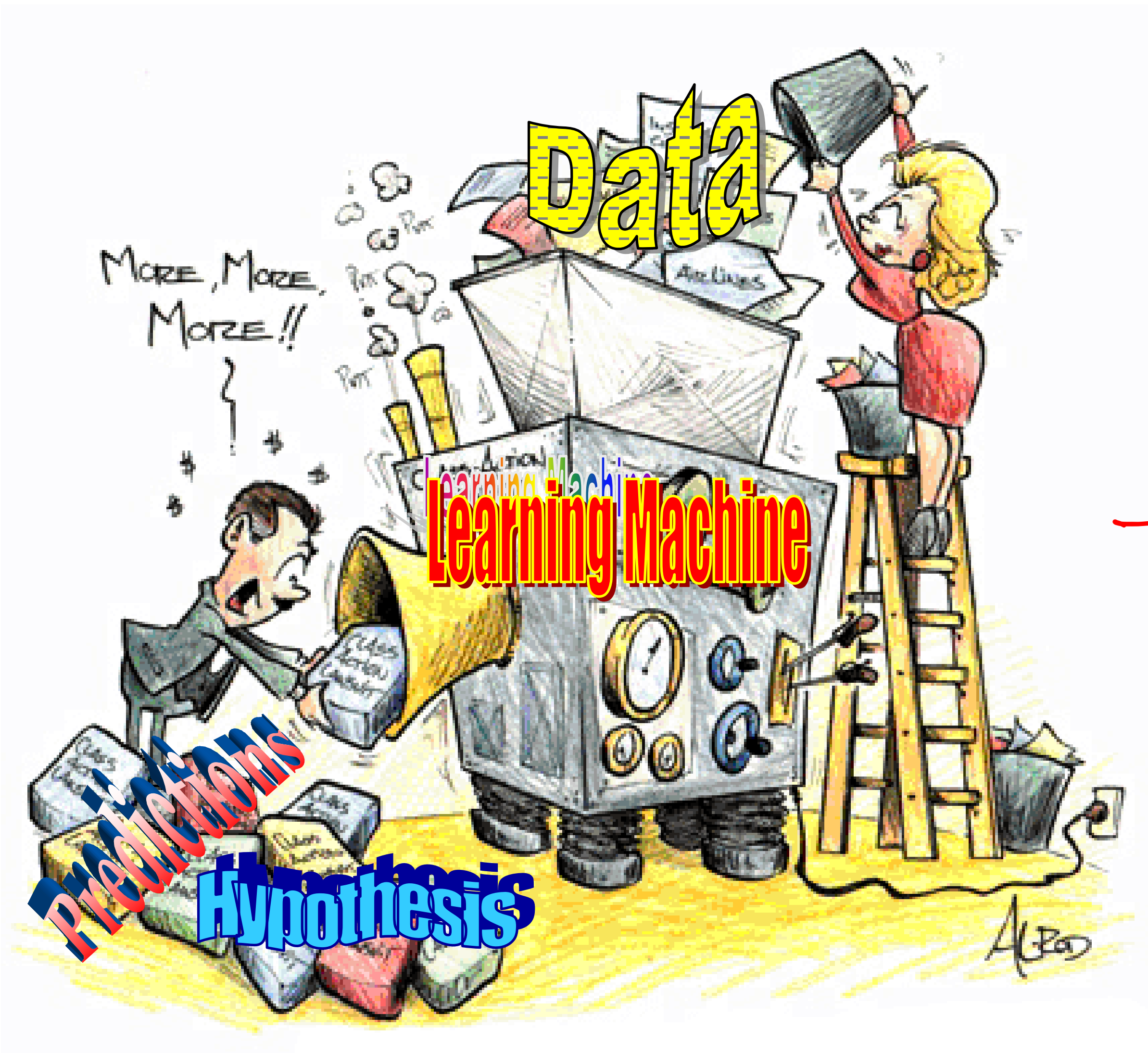
ML is trending!

– Wide applicability
– Very large-scale complex systems
  • Internet (billions of nodes), sensor network (new multi-modal sensing devices), genetics (human genome)
– Huge multi-dimensional data sets
  • 1.6 million images, 1000 object categories
  • 30,000 genes x 10,000 drugs x 100 species x …
– Software too complex to write by hand
– Improved machine learning algorithms
– Improved data capture (Terabytes, Petabytes of data), networking, faster computers
– Demand for self-customization to user, environment

“Data scientist: The sexiest job of the 21st century”

(Harvard Business Review)

What is Machine Learning?



Data

⇓

Learning algorithm

⇓

Knowledge

What is Machine Learning?

Design and Analysis of algorithms that

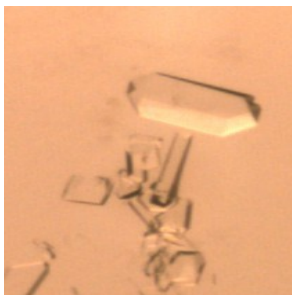
- improve their performance
- at some task
- with experience

(experience) **(performance on task)**

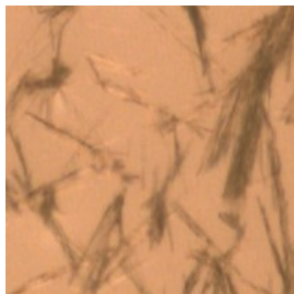
http://phillips-lab.biochem.wisc.edu/

Human learning

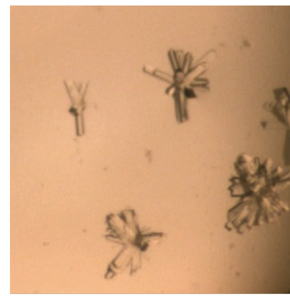
Task: Learning stage of protein crystallization



Crystal

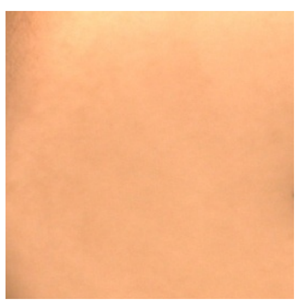


Needle

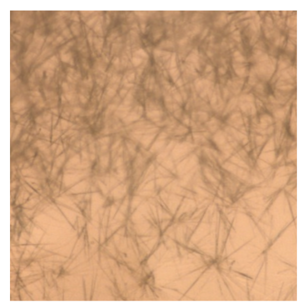


Tree

Tree



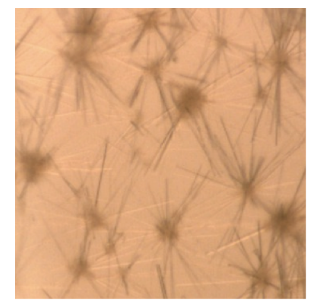
Empty



Needle

Experience

- Predict the label of the test image?



?

Performance

Tasks, Experience, Performance

Tasks, Experience, Performance

Machine Learning Tasks

Broad categories -

- **Supervised learning**

  Classification, Regression

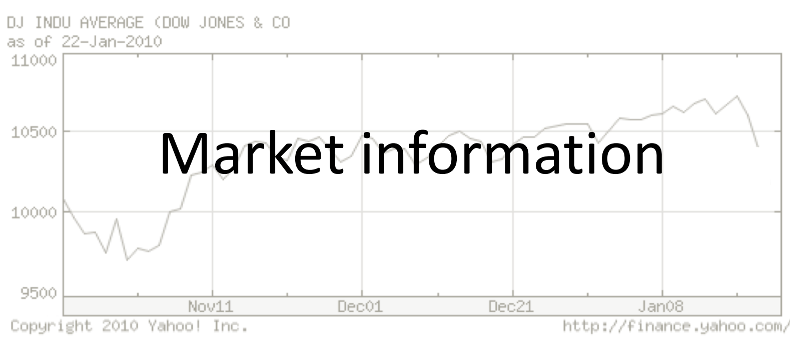
- **Unsupervised learning**

  Density estimation, Clustering, Dimensionality reduction

- **Graphical models, Hidden Markov models**
- **Reinforcement learning**
- Semi-supervised learning
- Active learning
- Many more …

Supervised Learning

Input $X \in \mathcal{X}$



Label $Y \in \mathcal{Y}$

"Sports"
"News"
"Science"
...

Discrete Labels
Classification

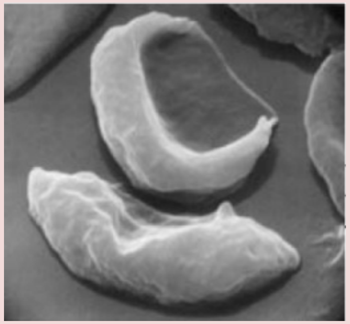
Share Price
"$ 24.50"

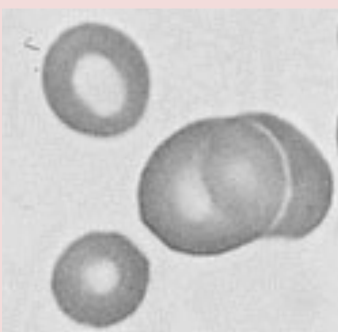
Continuous Labels
Regression

Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

$\equiv$ Construct **prediction rule** $f : \mathcal{X} \to \mathcal{Y}$

Classification or Regression?



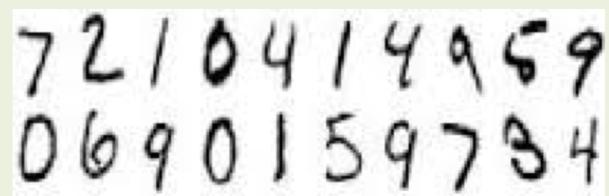


Medical Diagnosis

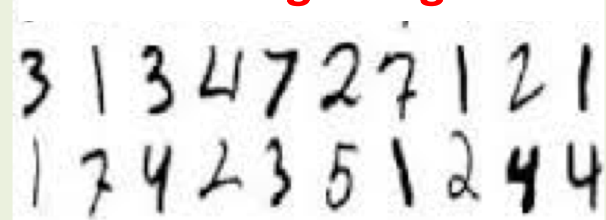
"Anemic"
"Healthy"

11 am	12 pm	1 pm	2 pm	3 pm	4 pm	5 pm	6 pm
39° F	41° F	44° F	44° F	44° F	44° F	43° F	42° F
Precip: 10%	Precip: 10%	Precip: 10%	Precip: 10%	Precip: 10%	Precip: 10%	Precip: 10%	Precip: 0%

Weather prediction



Handwriting recognition



Unsupervised Learning

Aka "learning without a teacher"

Input $X \in \mathcal{X}$

Word distribution
(Probability of a word)

$p(X=x)$

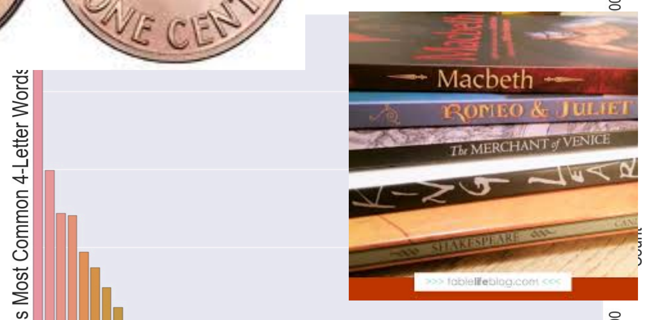
$\boxed{\textbf{Task:} \quad \text{Given } X \in \mathcal{X}, \text{ learn } f(X).}$

Unsupervised Learning

Learning a Distribution

Bias of a coin

Distribution of words in text



Img src: tablelifeblog.com ,data.world

Clustering

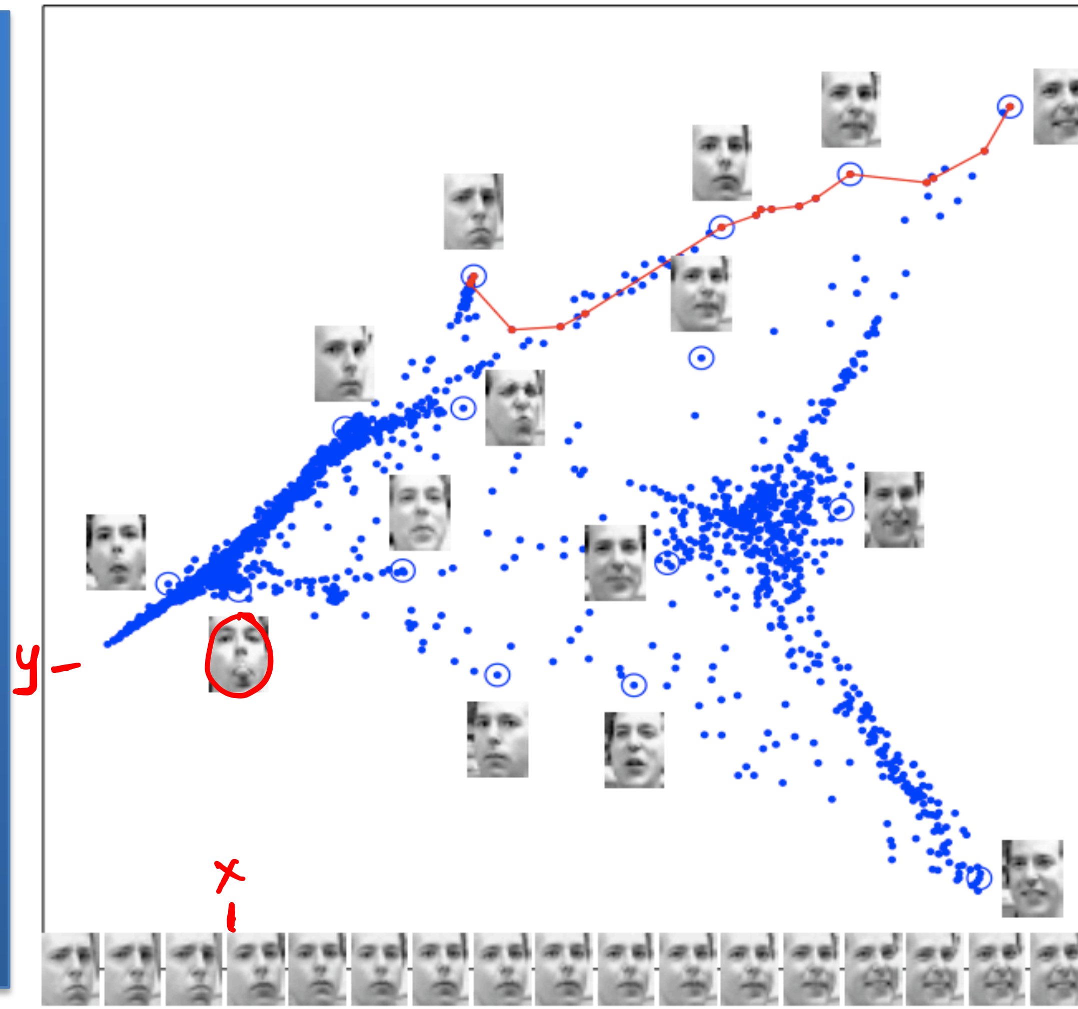
Unsupervised Learning

Dimensionality Reduction/Embedding

[Saul & Roweis '03]

Images have thousands or millions of pixels.

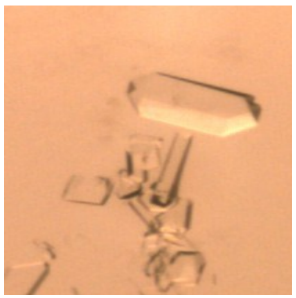
Can we give each image a small set of coordinates, such that similar images are near each other?

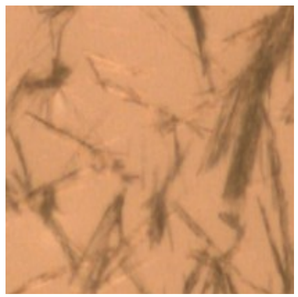


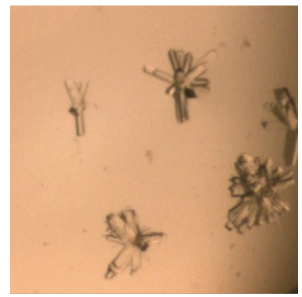
Tasks, **Experience**, Performance

Experience = Training Data

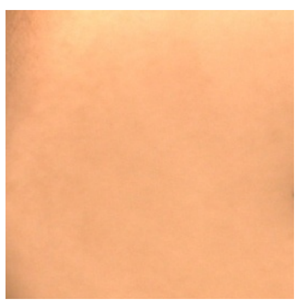
Task: Learning stage of protein crystallization

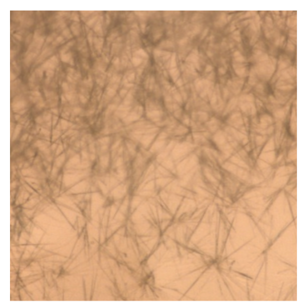

Crystal


Needle

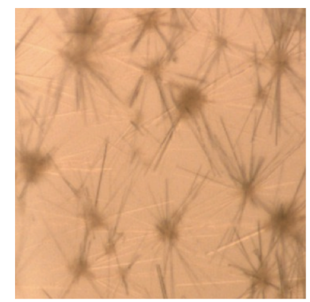

Tree


Tree


Empty


Needle

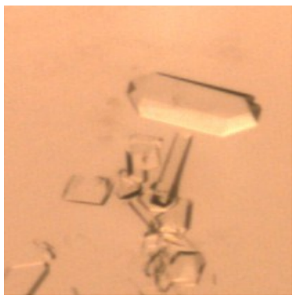
Experience


?

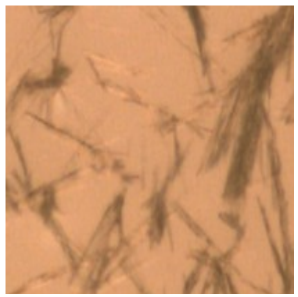
Performance

Training Data ≠ Test Data

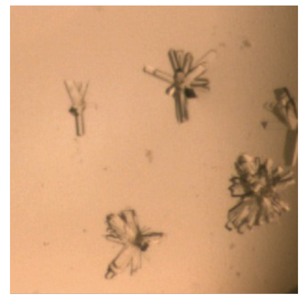
Task: Learning stage of protein crystallization



Crystal

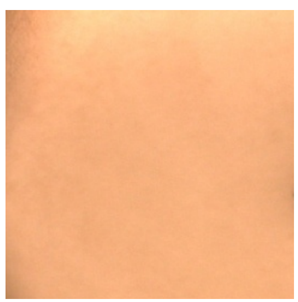


Needle

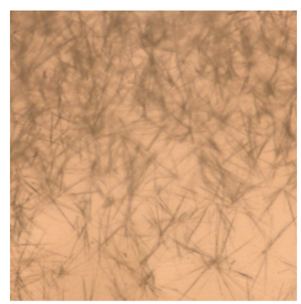


Tree

Tree

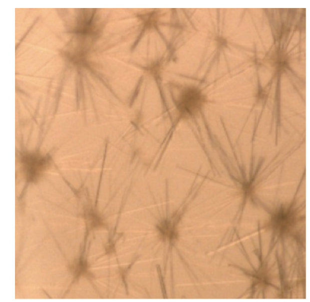


Empty



Needle

Experience



?

Performance

Critical to report testing and NOT training accuracy

Regression example: Blood samples were collected for 100 subjects who were administered a covid-19 vaccine.

An ML algorithm was trained to predict the number of antibodies in the blood of these 100 subjects given their profiles.

The normalized mean square error of the trained model was 0.001 for predicting the antibodies in these 100 subjects.

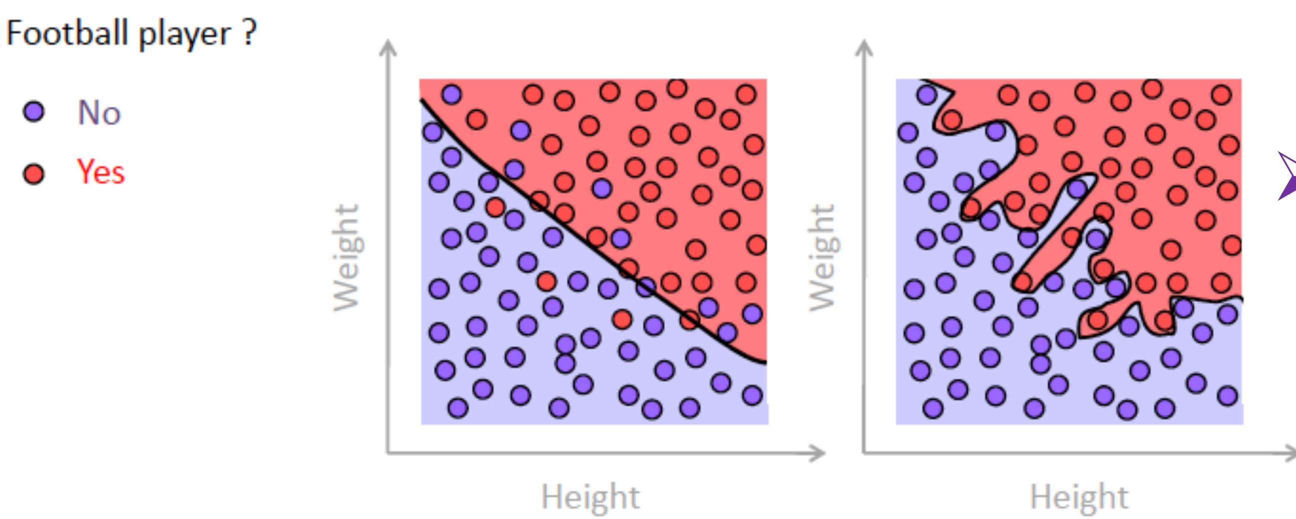
- Is this a good model?

10 more subjects were then recruited and the normalized mean square error of the model's predictions of antibodies for these 10 subjects was 0.35.

Generalization and overfitting

- What is learning really?

Can algorithm **generalize** aka perform well on test data



➢ Which is better (left/right)?

- Is it okay to **overfit** (get 0 error on) training data?

  Sometimes yes (e.g. if labels are noiseless)

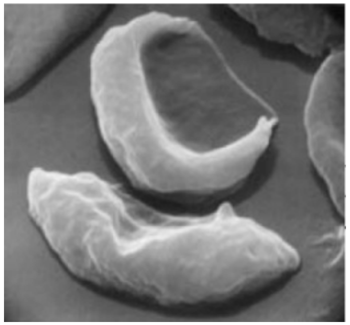
  BUT needs to be accompanied with ability to do well on unseen data

Tasks, Experience, Performance

Performance Measure

Performance:

$\text{loss}(Y, f(X))$ - Measure of closeness between label *Y* and prediction *f*(*X*) for test data X

X	Diagnosis, *Y*	*f*(*X*)	$\text{loss}(Y, f(X))$
	"Anemic cell"	"Anemic cell"	0
		"Healthy cell"	1

$$\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$$ **0/1 loss**

Performance Measure

Performance:

$\text{loss}(Y, f(X))$ - Measure of closeness between label *Y* and prediction *f*(*X*) for test data X

X	Share price, *Y*	*f*(*X*)	$\text{loss}(Y, f(X))$
Past performance, trade volume etc. as of Sept 8, 2010	"$24.50"	"$24.50"	0
		"$26.00"	1?
		"$26.10"	2?

$$\boxed{\text{loss}(Y, f(X)) = (f(X) - Y)^2 \quad \textbf{squared loss}}$$

Performance Measure

For test data X, measure of closeness between label *Y* and prediction *f*(*X*)

Binary Classification $\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$ **0/1 loss**

Regression $\text{loss}(Y, f(X)) = (f(X) - Y)^2$ **squared loss**

Lets think of unsupervised tasks next.

Performance Measure

For test data X, measure how good is the learnt distribution, clustering or embedding $f(X)$

Learning a distribution

- What performance measure would you use?

Glossary of Machine Learning

- Task
- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Learning distribution
  - Clustering
  - Dimensionality reduction/Embedding
- Input, X
- Label, Y
- Prediction, f(X)
- Experience = Training data
- Test data
- Overfitting
- Generalization
- Performance
- Likelihood
- Loss – 0/1, squared, negative log likelihood